现代汉语句型自动识别的研究

李伟

指导教师 曾文华教授

厦门大学

厦门大学博硕士论文摘要库

学校编码：10384　　分类号_____密级 _____

学号：200440023　　UDC _____

厦门大学

硕 士 学 位 论 文

现代汉语句型自动识别的研究

Research on Automatic Recognition of Sentence Patterns of Modern Chinese

李 伟

指导教师姓名：　曾 文 华 教授

专 业 名 称：　计算机应用技术

论文提交日期：　2007 年 5 月

论文答辩日期：　2007 年　　月

学位授予日期：　2007 年　　月

答辩委员会主席：_____

评　　阅　　人：_____

2007 年 5 月

厦门大学博硕士论文摘要库

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文产生的权利和责任。

声明人（签名）：

年　　月　　日

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1、保密（　），在　　年解密后适用本授权书。

2、不保密（√）

（请在以上相应括号内打“√”）

作者签名：　　日期：　　年　月　日

导师签名：　　日期：　　年　月　日

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

摘　要

每一种语言都有自身的特点，一种语言区别另一种语言，句型上的差异起着重要作用。汉语句型复杂多样，是汉语句法理论中一个重要的研究单位。但是，在中文信息处理中，以句型为目标的研究并不多。本文以汉语句型的自动识别为研究对象，强调结合自然语言处理要求和汉语语法规律，构建统一的句型系统框架，并在该框架下，尝试进行句型分析、研究句型识别策略。

在汉语句型的语言理论研究中，本文结合句模理论，提出“阶”的概念，构建了将谓词的句法特性与语义特征结合的谓词知识库，为计算机识别句型提供帮助。

在汉语句型识别的策略研究中，本文给出了汉语句型自动识别系统的流程框架。该框架中包含预处理模块和句型识别模块两个部分。预处理模块以去除句子的非句型成分为目的，抽象出句子的句干，填充包括定中结构、状中结构和补中结构的句法关系槽。句型识别模块以判定句干的句型类别为主要目的，并填充包含主谓关系、谓宾关系的句法关系槽。

在预处理模块中，本文提出“语片”的概念，将预处理过程分为粘合语片、填充句法关系槽两个流程，并提出基于滑动窗口机制的粘合算法，以及基于上下文无关文法的填充器 C-Filler。在句型识别模块中，本文提出了“基于规则匹配”和“基于函数模型”的两种可选策略。“基于规则匹配”的策略中，利用扩充的上下文无关文法，设计了句型识别器 Recognizer；“基于函数模型”的策略中，本文首次提出，利用转化函数，将句子转化到坐标空间中，使用函数模型研究句子的句法特征，并通过函数计算获得句子的特征向量用于机器学习，最终利用机器学习得到的决策函数识别句型。“基于函数模型”的策略是一种“引用数学方法研究句子”的新思路。

作为策略评估，在句型系统的子集上构建了实验模型，对预处理策略和两种句型识别策略分别进行了评估，实验结果证明了三种策略的可行性。

最后，论文从应用角度，在机器翻译、语法错误自动检查、统计汉语句型分布等几个方面，对汉语句型自动识别的应用前景作了简介。

厦门大学博硕士论文摘要库

关键词：汉语句型；谓词知识库；语片；基于规则匹配的识别；基于函数模型的识别

厦门大学博硕士论文摘要库

Abstract

Each language has its own characteristics, which are different from one language to another one. During these differences, the difference on Sentence Patterns is an important one. Chinese Sentence Patterns are complex and diverse, which constitute the most important parts of Chinese syntactic theory. However, in the Chinese Information Processing, the study on Sentence Patterns is not too much. This paper focuses on automatic recognition of Sentence Patterns of modern Chinese and emphasizes building a unified system framework on Sentence Patterns, with combining the rules of Chinese grammar and the requirements of Natural Language Processing (NLP). In the unified system framework, we try to analyze the Sentence Patterns and do research on the strategies of recognition of Sentence Patterns.

In the theoretical study of Chinese Sentence Patterns, this paper presents a new concept "Bands", learning from the theory of Chinese Sentence Mode which says semantic features of sentence, and constructs Predicate Knowledge Base which describes both the syntactic and the semantic features of predicates. This Predicate Knowledge Base provides convenience for computer to recognize the Sentence Patterns.

In the strategy study of the recognition of Chinese Sentence Patterns, this paper presents the framework of automatic recognition system. This framework includes two parts, the pre-processing module and the recognition module. The pre-processing module abstracts the sentences' stems, fills the troughs of syntactic relations, for the purpose of removing the non-Sentence Patterns ingredients. The recognition module mainly aims at determining which Sentence Patterns the stems belong.

厦门大学博硕士论文摘要库

In the pre-processing module, a concept "Fragment" is presented. This paper divides this module into two steps, agglutinating fragments and filling the troughs of syntactic relations, proposes agglutination algorithm based on the mechanism of the sliding window, and designs the algorithm "C-Filler" based on the context-free grammar (CFG). In the recognition module, this paper presents two optional strategies, Rule-based Strategy (RBS) and Function-based Strategy (FBS). In the RBS, based on the enhanced CFG, algorithm "Recognizer" for recognizing the Sentence Patterns is designed. In the FBS, it is proposed for the first time that transforms the sentence stem into a series of data in the coordinate space by transform function, gets eigenvector by function fitting, and finally gains the decision functions by Machine Leaning. FBS is a new idea on how to analyze sentences with mathematic methods.

To evaluate the strategies proposed, an experimental model is built. Pre-processing strategy, RBS strategy, and FBS strategy are evaluated and the results prove the feasibility of these three strategies.

Finally, several applications of automatic recognition system of Sentence Patterns, including Machine Translation, Grammatical Mistakes Automated Checks, and Analyzing Distribution of Chinese Sentence Patterns, are introduced.

Keywords: Chinese Sentence Patterns; Predicate Knowledge Base; Fragment; Rule Based Strategy (RBS); Function Based Strategy (FBS).

厦门大学博硕士论文摘要库

目　录

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

Contents

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

Degree papers are in the “Xiamen University Electronic Theses and Dissertations Database”. Full texts are available in the following ways:
1. If your library is a CALIS member libraries, please log on http://etd.calis.edu.cn/ and submit requests online, or consult the interlibrary loan department in your library.
2. For users of non-CALIS member libraries, please mail to etd@xmu.edu.cn for delivery details.

厦门大学博硕士论文摘要库